

# Network Discovery by Generalized Random Walks

A. Asztalos<sup>1,2</sup> and Z. Toroczkai<sup>1\*</sup>

<sup>1</sup> Interdisciplinary Center for Network Science and Applications (iCeNSA) and  
Department of Physics, University of Notre Dame, Notre Dame, IN, 46556 USA

<sup>2</sup>Department of Computer Science and Department of Physics,  
Rensselaer Polytechnic Polytechnic Institute, Troy, NY 12180-3590, USA

August 31, 2010

## Abstract

We investigate network exploration by random walks defined via stationary and adaptive transition probabilities on large graphs. We derive an exact formula valid for arbitrary graphs and arbitrary walks with stationary transition probabilities (STP), for the average number of discovered edges as function of time. We show that for STP walks site and edge exploration obey the same scaling  $\sim n^\lambda$  as function of time  $n$ . Therefore, edge exploration on graphs with many loops is always lagging compared to site exploration, the revealed graph being sparse until almost all nodes have been discovered. We then introduce the Edge Explorer Model, which presents a novel class of adaptive walks, that perform faithful network discovery even on dense networks.

## 1 Introduction

Random walk theory [1, 2, 3] has seen myriad applications ranging from physics, biology and ecology through market models, finance, to problems in mathematics and computer science. It is being used to sample distributions, compute volumes and solve convex optimization problems [4] and has played a key role in www search engines [5]. It provides a microscopic description for real-world transport processes on networks, such as information spread [6, 7] and disease propagation (epidemics [8, 9, 10]) and it can also be used to design network discovery/exploration tools [11]. Here we focus on the latter aspect.

The structure of real-world networks [12, 13] is organically evolving and frequently, either their size is simply prohibitive for measuring their topology (the WWW has  $\sim 2 \times 10^{10}$  nodes), or the nature of the network makes it difficult to gather global information (e.g., in some social networks). Such networks are best explored by ‘walkers’ stepping from a node to a neighbor node linked by an edge, collecting local information, which is then combined to produce subgraph

---

\*E-mail: toro@nd.edu

samples of the original graph. To fix notations, we denote by  $G(V, E)$  the graph on which the walk happens, where  $V$  is the set of  $N$  nodes,  $E$  is the set of  $M$  edges, and by  $p(s'|s; t)$  the single-step transition probability of a walker at site (node)  $s$  to step onto a neighboring site  $s'$  ( $(s, s') \in E$ ), on the  $t$ -th step. Note that equivalently, one could consider the walk taking place on the complete graph  $K_N$ , with setting  $p(s'|s; t) = 0$  for  $(s', s) \notin E$ . One can think of  $p(s'|s; t)$  as information ‘handed’ to the walker at node  $s$  to follow in its choice for stepping to the next node. Accordingly, an important optimization problem is to ‘design’ the  $p(s'|s; t)$  probabilities such that certain properties of the exploration are optimal. Such problems motivate the development of the statistical mechanics of network discovery, connecting the set of local transition probabilities  $\{p(s'|s; t)\}$  with the global properties of the uncovered subgraph as function of time. We distinguish two main classes of exploration problems, namely those with: I. stationary transition probabilities (STP) where  $p(s'|s; t) = p(s'|s)$  (time-independent) and II. adaptive transition probabilities (ATP) where  $p(s'|s; t)$  depends on time and possibly on the past history of the walk. For general STP walks, analytic results were obtained for the number  $S_n$  of distinct (virgin) nodes visited in  $n$  steps (site exploration) [14, 15, 16] (for a review see [1]), and on the cover time  $T_V^*$  (expected number of steps to visit all nodes, for a review see [17]). Numerically, site exploration by simple random walks  $p(s'|s) = k_s^{-1}$  ( $k_s$  is the degree of node  $s$ ) has been extensively studied on various complex network models [18, 19].

Interestingly, the number  $X_n$  of distinct *edges* (edge exploration) visited in  $n$ -steps has only been studied numerically, for simple random walks, [20, 21], and no analytic results similar to  $S_n$  have been derived. The statistics of  $X_n$ , however, cannot be obtained directly from the analytic results for  $S_n$  on a “dual” graph, such as the edge-to-vertex dual graph  $L(G)$ , because the walk does not transform simply onto  $L(G)$ . On graphs with loops, there is an inherent asymmetry between the evolution of  $S_n$  and  $X_n$ . While a new node is always discovered via a new edge ( $S_{n+1} = S_n + 1$  implies  $X_{n+1} = X_n + 1$ , Fig. (1a)), a new edge can be discovered between two previously visited nodes as well (Fig. (1b)). In the latter case, the walker always encloses a loop in the discovered subgraph, hence the  $(S_n, X_n)$  pair can be connected to the *loop statistics* of the network. More precisely, the quantity  $Q_n = 1 + X_n - S_n$  gives the number of times the walker returned to its own path through a freshly discovered edge, in  $n$  steps. Clearly, if  $G$  is a tree, then  $X_n = S_n - 1$  for all  $n \geq 0$ .

In this Letter we provide an exact expression for the generating function for the average number of discovered edges  $\langle X_n \rangle$  in  $n$ -steps for arbitrary STP walks on arbitrary graphs. Although our expressions are valid in general, we are interested in the *scaling behavior* of  $\langle S_n \rangle$  and  $\langle X_n \rangle$  for large times ( $n \gg 1$ ) on large ( $N \gg 1$ ) *connected* graphs. Let us consider a monotonically increasing sequence  $\{c_n\}$  of positive terms  $c_n > 0$ . We will use the notation  $c_n \sim n^\kappa$  to say that  $c_n$  scales with  $n$  with a growth exponent  $\kappa$ , if as  $n \rightarrow \infty$ ,  $c_n \simeq n^\kappa L(n)$ , where  $L(n)$  is a slowly varying function that is  $L(\eta x)/L(x) \rightarrow 1$  as  $x \rightarrow \infty$  for any  $\eta > 0$ .

As it will be seen, for STP walks on large but finite graphs both the average number of discovered nodes and edges obey scaling laws  $\langle S_n \rangle \sim n^\lambda$ ,  $\langle X_n \rangle \sim n^\mu$ . These hold up to a *cross-over time*  $T_V$  (for  $\langle S_n \rangle$ ) and  $T_E$  (for  $\langle X_n \rangle$ ) after which *saturation* sets in until all the nodes (edges) have been discovered, at the corresponding cover times  $T_V^*$  and  $T_E^*$ . At the cross-over time only a small *constant number* of nodes (edges) are left untouched and we consider this as the time where the discovery has practically been completed. Since in a step at most one node (edge) can be

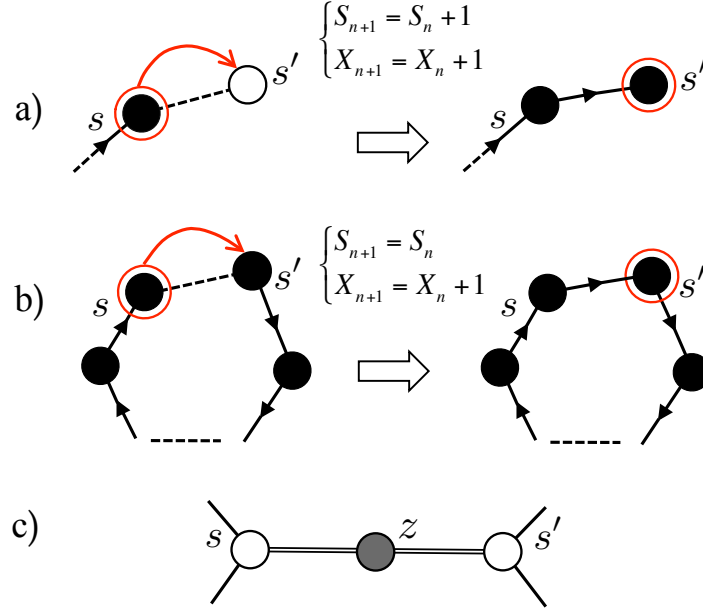


Figure 1: a) A new node (empty circle) is always discovered via a new edge (dashed line). b) A new edge, can also be discovered between already visited nodes. c) First passage time through an edge can be computed from the first passage time to a pseudonode  $z$  placed on the edge.

discovered, the growth of  $S_n$  ( $X_n$ ) is *at most linear*, at any time, for any walk (STP or ATP), that is  $\lambda \leq 1$ ,  $\mu \leq 1$ . When the walker is in completely charted territory, both  $S_n$  and  $X_n$  stagnate, otherwise  $X_n$  always grows (Fig. (1)a-b) ), and hence  $\mu \geq \lambda$ .

Here we show that for *recurrent* STP walks both site and edge exploration obey the same scaling that is  $\lambda = \mu$  in the  $N \rightarrow \infty$  limit. This means that in dense graphs where the nr of edges  $M \sim N^\nu$ ,  $\nu > 1$  (but  $\nu \leq 2$ ), the node set  $V$  is discovered much earlier than the edge set  $E$  ( $T_V < T_E$ ). As we prove below, even for the complete graph on  $N$  nodes  $G = K_N$ , STP walks explore nodes and edges at the same rate,  $\mu = \lambda$ . This is counterintuitive, because there are  $\mathcal{O}(N^2)$  edges, and the walker could keep discovering many new edges between already visited nodes, so there is no obvious reason why we could not have  $\mu > \lambda$ . The fact that for STP walks, edge and site exploration grow at the same rate, presents a problem if one is interested in discovering the *links* (relationships) in a network. Moreover, if network discovery is done with the purpose of sampling and producing a subgraph with statistical properties resembling that of the underlying network, then STP walks will not provide the optimal solution, independently on the form of the transfer matrix  $p(s|s')$ . This is simply because a walker's choice to move to a neighbor will be independent of its visiting history, and therefore will have a lower chance on average to discover a virgin edge to a visited neighbor (Fig. 1b)) than for e.g., an ATP walk that is biased towards already visited neighbors. Hence, for a given number of visited nodes in an STP walk  $\langle X_n \rangle = \mathcal{O}(\langle S_n \rangle)$  number of edges will be revealed before  $T_V$ , making the discovered subgraph sparse, seriously skewing the sample especially, if the underlying network is dense ( $\nu > 1$ ). To

resolve this, we introduce an ATP walk, the Edge Explorer Model (EEM) that performs a faithful exploration of the nodes and edges even on dense networks, such that  $T_E \simeq T_V$ .

## 2 STP walks

The generating function  $S(s_0; \xi) = \sum_{n=0}^{\infty} \langle S_n \rangle \xi^n$  for the average number of distinct sites discovered in  $n$  steps by the walker starting from site  $s_0$  can be written as [1]:

$$S(s_0; \xi) = \frac{1}{1-\xi} \sum_{s \in V} W(s; \xi) P(s|s_0; \xi), \quad (1)$$

$$W(s; \xi) = [P(s|s; \xi)]^{-1}, \quad (2)$$

where  $P(s|s_0; \xi)$  is the site occupation probability generating function, that is  $P(s|s_0; \xi) = \sum_{n=0}^{\infty} \xi^n P_n(s|s_0)$ , with  $p_n(s|s_0)$  being the probability for a walker starting from  $s_0$  to be found at  $s$  on the  $n$ -th step. Next we derive a similar expression for  $X(s_0; \xi) = \sum_{n=0}^{\infty} \langle X_n \rangle \xi^n$ . Let  $F_n(s|s_0)$  be the first-passage time distribution (the probability for the walker to arrive at  $s$  for the first time on the  $n$ -th step) and let  $F(s|s_0; \xi)$  be its generating function. It is well known [1] that  $F(s|s_0; \xi) = [P(s|s_0; \xi) - \delta_{s,s_0}] / P(s|s; \xi)$ . The probability that  $s$  is ever reached by the walker starting from  $s_0$  is therefore  $R(s|s_0) = \sum_{n=1}^{\infty} F_n(s|s_0) = \lim_{\xi \rightarrow 1, |\xi| < 1} F(s|s_0; \xi)$ . Since  $R(s|s_0) \leq 1$ ,  $F(s|s_0; 1^-)$  is convergent, however,  $P(s|s_0; 1^-)$  can diverge. If  $R(s|s_0) = 1$  for all  $s, s_0$ , then the walk is *recurrent*,  $P(s|s_0; 1^-) = \infty$  and  $F(s|s_0; 1^-) = 1$ . Moreover, in this case  $P(s|s_0; 1^-)$  has the *same rate of divergence* for all  $s, s_0 \in V$  [1]. For finite networks, in which the walker can access all nodes and there are no traps, the walk is recurrent.

Let  $F_n(e|s_0)$  denote the *edge* first-passage time distribution, i.e., the probability that the walker passes through edge  $e = (s, s') \in E$  for the first time on the  $n$ -th step, given that it started at node  $s_0$ . By introducing an indicator  $\Gamma_n(s_0)$  for the number of virgin edges discovered on the  $n$ th step ( $= 0, 1$ ), we have  $\langle \Gamma_n \rangle = \text{Prob}\{\Gamma_n = 1\} = \sum_{e \in E} F_n(e|s_0)$ , with  $\Gamma_0 = \langle \Gamma_0 \rangle = 0$ . Clearly,  $X_n(s_0) = \sum_{j=0}^n \Gamma_j(s_0)$  and thus the generating function for the average number of visited distinct edges in  $n$  steps becomes:

$$X(s_0; \xi) = \frac{\Gamma(s_0; \xi)}{1-\xi} = \frac{1}{1-\xi} \sum_{e \in E} F(e|s_0; \xi), \quad (3)$$

where  $\Gamma(s_0; \xi)$  and  $F(e|s_0; \xi)$  are generating functions for  $\Gamma_n(s_0)$  and  $F_n(e|s_0)$ , respectively.

To obtain the edge first-passage time distribution, we place an auxiliary site  $z$  on edge  $e = (s, s') \in E$  (Fig 1c)) and redefine the walk on this new graph  $G_z$  such that the *node* first-passage time probability to  $z$  on  $G_z$  is the same as the *edge* first-passage time probability through  $e$  on  $G$ . The extended graph  $G_z(V_z, E_z)$  has  $V_z = V \cup \{z\}$  and  $E_z = \{(s, z); (z, s')\} \cup E \setminus \{(s, s')\}$ . The addition of  $z$  to  $e = (s, s')$  changes only the transition probabilities around that edge, leaving  $p(r|r')$  the same away from  $s$  and  $s'$ . Steps from  $s$  ( $s'$ ) to  $s'$  ( $s$ ) in the new walk are forbidden; instead, the walker has to step onto node  $z$  first. However, the same probability flow has to exist in the modified walk as in the original one when moving from sites  $s$  and  $s'$  towards  $z$ . From  $z$  the walker is only allowed to step to  $s$  or  $s'$  with arbitrary probabilities  $f$  and  $g$ , respectively

(which, however, should not enter the final expression for  $F(e|s_0; \xi)$ !). The single-step transition probabilities on the  $G_z$  for the new walk can thus be combined into  $(r, r' \in V_z)$ :

$$p^\dagger(r|r') = (1 - \delta_{r'z} - \delta_{rz} - \delta_{r's'}\delta_{rs} - \delta_{r's}\delta_{rs'})p(r|r') + \delta_{rz}\delta_{r's'}p(s|s') + \delta_{rz}\delta_{r's}p(s'|s) + \delta_{r'z}(f\delta_{rs} + g\delta_{rs'}) . \quad (4)$$

The rest of the calculation focuses on obtaining the node first-passage time distributions  $F_n^\dagger(z|s_1)$  and site occupation probabilities  $P_n^\dagger(r|s_1)$  of the modified walk  $(r, s_1 \in V_z)$ . Due to our setup we have  $F_n(e|s_0) = F_n^\dagger(z|s_0)$ ,  $s_0 \in V$ , or  $F(e|s_0; \xi) = F^\dagger(z|s_0; \xi) = P^\dagger(z|s_0; \xi)/P^\dagger(z|z; \xi)$ ,  $s_0 \neq z$ . Obtaining the  $P^\dagger$  generating functions in terms of the original functions  $P$  involves a lengthy series of Green-function manipulations, using the formalism developed for ‘taboo sites’ [1, 22], see Appendix A. The final result after using (3) is:

$$X(s_0; \xi) = \frac{\xi}{1 - \xi} \sum_{s \in V} \bar{W}(s; \xi) P(s|s_0; \xi) , \quad (5)$$

with

$$\bar{W}(s; \xi) = \sum_{s' \in V} \alpha \frac{1 + (d - c)\beta\xi}{1 + (a\alpha + d\beta)\xi + (ad - bc)\alpha\beta\xi^2} , \quad (6)$$

where  $a = a(\xi) = P(s|s'; \xi)$ ,  $b = b(\xi) = P(s'|s; \xi)$ ,  $c = c(\xi) = P(s'|s'; \xi)$ ,  $d = d(\xi) = P(s'|s; \xi)$  and  $\alpha = p(s'|s)$ ,  $\beta = p(s|s')$ . The form (5) is similar to (1), however with a more involved weight function  $\xi\bar{W}(s; \xi)$ . This shows that edge exploration is usually quite different from node exploration. Expressions (5-6) are entirely general, valid for any type of STP walk (including asymmetric walks  $p(s|s') \neq p(s'|s)$ ), on arbitrary graphs. The properties of  $\xi b(\xi)\bar{W}(s; \xi)$  fully determine the statistics of edge exploration, *when compared* to site exploration. Note that the summation in the expression of the weight  $\bar{W}(s; \xi)$  is only over the network neighbors of  $s$ , due to the multiplicative transition probability  $\alpha = p(s'|s)$ , which is zero if  $s'$  and  $s$  are not neighbors in the graph. Next, we discuss some special cases for simple random walks.

### 3 Special cases

*I. Simple random walks on  $K_N$ .* The single-step probabilities of a simple random walker can be written as  $p(s|s') = p(1 - \delta_{ss'})$ ,  $p = 1/(N - 1)$ . In this case

$$P(s|s_0; \xi) = [\delta_{s,s_0} + p\xi(1 - \xi)^{-1}] (1 + p\xi)^{-1} ,$$

and  $S(\xi)$  and  $X(\xi)$  are easily obtained. In particular,

$$S(\xi) = (1 - \xi)^{-1}(1 + p\xi) [1 - (1 - p)\xi]^{-1} ,$$

from where, via contour integration,  $\langle S_n \rangle = [1 + p - (1 - p)^n] / p$ . Similarly,

$$\bar{W}(s; \xi) = [1 + (a + b)p\xi]^{-1} ,$$

and since in this case  $\xi a = b - 1$ , we have

$$\overline{W}(s; \xi) = [1 - p + (1 + \xi)pb]^{-1}.$$

From (5) and the expression for  $b = P(s|s; \xi)$  it follows:

$$X(\xi) = \frac{\xi}{1 - \xi} \cdot \frac{1 - \xi + Np\xi}{1 + (2p - 1)\xi + 2p(p - 1)\xi^2}. \quad (7)$$

After contour integration we obtain the exact expression

$$\langle X_n \rangle = \frac{1 + p}{2p^2} + \frac{q_1 q_2}{q_1 - q_2} \left[ \frac{1 + pq_1}{q_1^n (q_1 - 1)} - \frac{1 + pq_2}{q_2^n (q_2 - 1)} \right], \quad (8)$$

$n \geq 0$ . Here  $q_1 > 1$ ,  $q_2 < -1$  are the roots of the quadratic equation  $2p(1 - p)\xi^2 - (1 - 2p)\xi - 1 = 0$ . Fig. (2)a shows the agreement between simulations and the analytical formulas for  $\langle S_n \rangle$ ,  $\langle X_n \rangle$ . From the above, for large graphs,  $p \ll 1$ ,  $\langle S_n \rangle - 1 = p^{-1} [1 - (1 - p)^n] = n - \binom{n}{2}p + \dots$ , showing that  $\langle S_n \rangle \sim n$  in the regime  $np \ll 1$ , or  $n \ll N$ . Similarly, for large graphs,  $q_1 = 1 + 2p^2 + \mathcal{O}(p^3)$ ,  $q_2 = -\frac{1}{2p} - \frac{1}{2} + \mathcal{O}(p)$ , yielding  $\langle X_n \rangle \sim (1 + p)(n - 1 + \dots)$  for  $n \ll N^2$ . The cover times can also be calculated, yielding  $T_V^* \sim N \ln N$  (coinciding with [17]) and  $T_E^* \sim N^2 \ln N$ . Thus, both  $\langle S_n \rangle$  and  $\langle X_n \rangle$  grow *together*, linearly, ( $\lambda = \mu = 1$ ) as function of time, with  $\langle S_n \rangle$  saturating to  $N$  at  $T_V$  and  $\langle X_n \rangle$  to  $N(N - 1)/2$  much later, at  $T_E$ . Fig. (2)b shows the linear dependence of  $\langle X_n \rangle$  on  $\langle S_n \rangle$ .

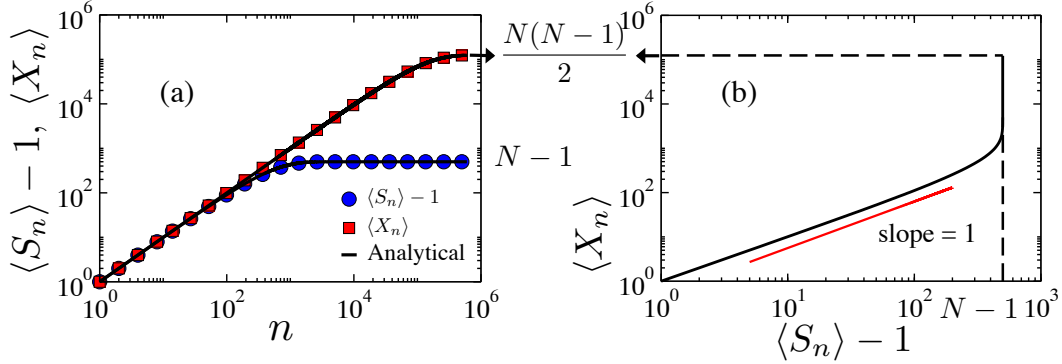


Figure 2: (a) Numerical (blue circles for  $\langle S_n \rangle$ , red squares for  $\langle X_n \rangle$ , respectively) and analytical (black solid line) results shown for a complete graph of 500 nodes. (b) Same as a) with  $\langle X_n \rangle$  vs  $\langle S_n \rangle$ , showing that the discovered graph is sparse even on  $K_N$ .  $\langle S_n \rangle - 1$  is plotted instead of  $\langle S_n \rangle$  in order to account for the initial conditions  $S_0 = 1$ ,  $X_0 = 0$ .

*II. Infinite translationally invariant lattices in  $d$  dimensions.* Simple random walks on such graphs are homogeneous, that is  $p(s|s') \equiv p(\mathbf{l})$ ,  $P(s|s'; \xi) \equiv P(\mathbf{l}; \xi)$ , where  $\mathbf{l} = s - s'$ . It is known that  $P(\mathbf{l}; \xi) = (2\pi)^{-d} \int_B d\mathbf{k} e^{-i\mathbf{k}\mathbf{l}} [1 - \xi\omega(\mathbf{k})]^{-1}$  where the integration is over the first Brillouin zone  $B = [-\pi, \pi]^d$  and  $\omega(\mathbf{k}) = \sum_{\mathbf{l}} p(\mathbf{l}) e^{i\mathbf{k}\mathbf{l}}$  is the structure function of the walk. While exact formulas are hard to obtain in  $d \geq 4$ , the leading order behavior of  $\langle S_n \rangle$  and  $\langle X_n \rangle$  can

be extracted from applying the discrete Tauberian theorem [1] on the corresponding generating functions. According to this theorem, the scaling  $c_n \sim n^\kappa$  as  $n \gg 1$  is equivalent to having the behavior  $C(\xi) \simeq (1 - \xi)^{-\kappa-1} L(1/(1 - \xi))$  for the generating function  $C(\xi) = \sum_{n=0}^{\infty} \xi^n c_n$  in the limit  $\xi \rightarrow 1^-$ . The results for  $\langle S_n \rangle$  are also summarized in [1], we quote them here along with our results for  $\langle X_n \rangle$  for comparison and completeness. For  $d = 1$ ,  $\langle S_n \rangle \sim \sqrt{8n/\pi}$ ,  $\langle X_n \rangle \sim \sqrt{8n/\pi}$ . For  $d = 2$ , square lattice  $\langle S_n \rangle \sim \frac{\pi n}{\ln(8n)}$ ,  $\langle X_n \rangle \sim \frac{4\pi n}{3\pi + 2\ln(8n)}$  and for the triangular lattice  $\langle S_n \rangle \sim \frac{2\pi n}{\sqrt{3}\ln(12n)}$ ,  $\langle X_n \rangle \sim \frac{6\pi n}{5\pi + \sqrt{3}\ln(12n)}$ . For  $d \geq 3$  cubic (hypercubic) lattices, simple random walks are non-recurrent (transient), hence  $P(0; 1^-) < \infty$  and we obtain  $\langle S_n \rangle \sim \frac{n}{P(0; 1^-)}$ ,  $\langle X_n \rangle \sim \frac{4dn}{2d-1+2P(0; 1^-)}$ .

These special cases suggest that for simple random walks  $\lambda = \mu$ , i.e., the edges are discovered mostly by visiting new nodes, and once the nodes have all been visited, the remaining edges are discovered, at the same rate. This holds for simple random walks on other networks as well, as indicated by our simulations summarized in Fig. (3). Simulations were run on Erdős-Rényi (ER)

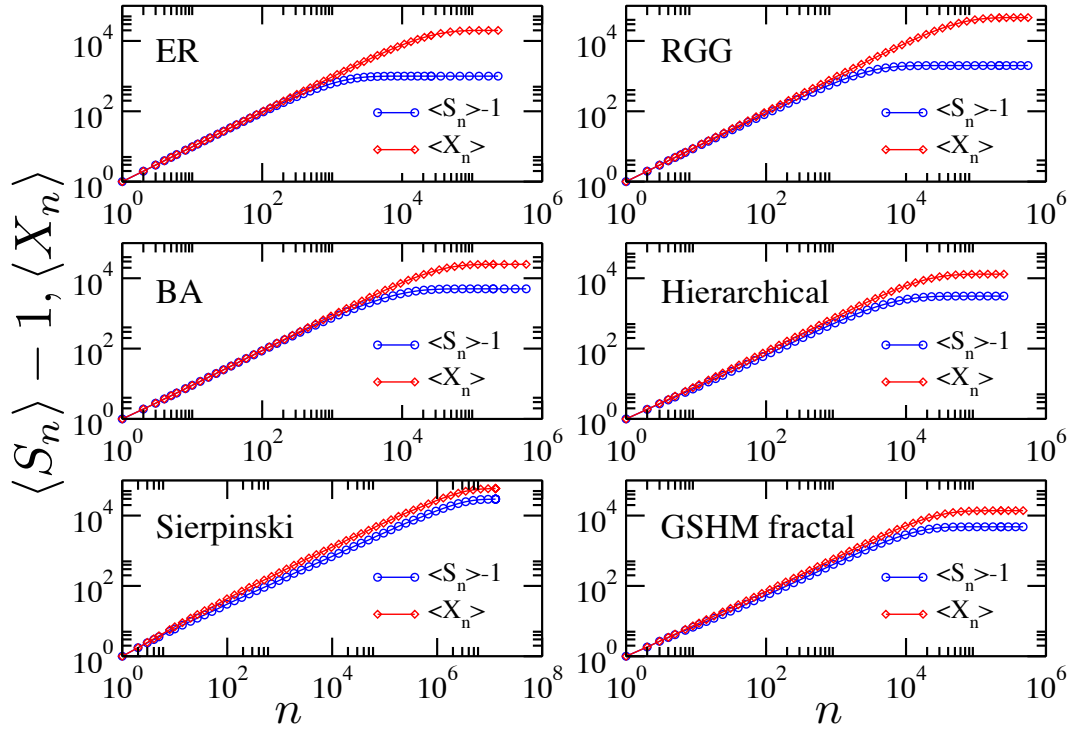


Figure 3: Growth of  $\langle S_n \rangle - 1$  (blue circles) and  $\langle X_n \rangle$  (red circles) for simple random walks. ER ( $N = 1000$ ,  $\langle k \rangle = 40$ ); RGG ( $d = 2$ ,  $N = 2000$ ,  $\langle k \rangle = 50$ ); BA ( $N = 5000$ ,  $\langle k \rangle = 10$ ); Hierarchical network ( $N = 3125$ ,  $\langle k \rangle = 8$ ); Sierpinski gasket ( $N = 29526$ ); GSHM fractal ( $N = 4810$ ,  $\langle k \rangle = 5.74$ ). Results were averaged for 300 initial conditions on 200 different graphs.

random graphs [23], random geometric graphs (RGG) [24] and the scale-free Barabási-Albert

(BA) model [25], the hierarchical network model [26], the fractal Sierpinski gasket and the GSHM fractal network [27]. The curves for  $\langle S_n \rangle$  and  $\langle X_n \rangle$  almost perfectly overlap, or run in parallel ( $\lambda = \mu$ ). In the case of ER, RGG and BA models the growth rates are linear. For the hierarchical network  $\lambda \simeq 0.94$ ,  $\mu \simeq 0.99$ , and for GSHM:  $\lambda \simeq 0.92$  and  $\mu \simeq 0.96$ . For the Sierpinski gasket  $\lambda \simeq 0.68$  (same as in [28]) and  $\mu \simeq 0.73$ . In all cases where deviations were observed for the exponents, they were small, on the order of  $\mu - \lambda \leq 0.05$ . One can show that these deviations are due to correction terms which, while vanish in the  $N \rightarrow \infty$  limit, they still show up in the simulations (which are on relatively small networks, to be able to observe the saturations). It is possible to prove that  $\lambda = \mu$  in the  $N \rightarrow \infty$  limit holds for general STP walks on arbitrary graphs, by showing that  $0 < b(1^-)\bar{W}(s; 1^-) < \infty$ , see Appendix B.

## 4 Adaptive walks: a simple bound

If STP walks are not good explorers, then naturally the question arises: What ATP walks would have good discovery properties? Due to time dependence, ATP walks present a much wider array of possibilities and their systematic treatment is a hard problem. Instead of tackling this general issue, here we first provide a simple *upper bound* for the *mean* edge discovery growth exponent, obeyed by *any* walk (ATP, or STP). Note that for ATP walks it is not necessarily true that  $\langle S_n \rangle$  or  $\langle X_n \rangle$  obeys scaling with a single exponent until saturation. However, due to the constraints from Fig. (1)a-b), the ‘local slopes’ still obey  $1 \geq d\langle X_n \rangle/dn \geq d\langle S_n \rangle/dn \geq 0$ . Because slopes vary, we define the mean growth exponents  $\mu = \ln M / \ln T_E$  and  $\lambda = \ln N / \ln T_V$ . The bound is based on the observation that the edge cross-over time cannot be smaller than the node cross-over time,  $T_E \geq T_V$ . This provides the upper bound  $\mu \leq \ln M / \ln(T_V)$ . At  $T_V$ , however,  $S_{n=T_V} \sim T_V^\lambda = N$ , and thus  $\ln(T_V) = \frac{1}{\lambda} \ln N$ . Recall that  $M \sim N^\nu$ . Since the graph  $G$  is connected,  $1 \leq \nu \leq 2$ . We therefore find that:

$$\lambda\nu \geq \mu \geq \lambda. \quad (9)$$

Hence a *necessary* condition for ATP walks to achieve  $\mu > \lambda$  mean growth exponents is  $\nu > 1$ . If  $\nu = 1$ , (sparse graphs), clearly no walk (ATP or STP) can achieve  $\mu > \lambda$ . This is for example the case for all large networks with  $N \rightarrow \infty$  that have fixed maximum degree  $D$ , as  $M \leq DN$  and thus  $\nu = 1$ . Inequalities (9) also show that the denser the graph, the larger the difference  $\mu - \lambda$  could be, possibly obtained by sufficiently ‘smart’ ATP walks. However, as  $\nu \leq 2$ , the mean edge discovery growth exponent can never be larger than twice that for nodes, i.e.,  $2\lambda$ .

## 5 The Edge Explorer Model (EEM)

Next we introduce an ATP walk, the Edge Explorer Model, where the transition probability to step onto a neighboring site depends on the visitation history of that site and its neighbors. The EEM is one of the simplest models that performs enhanced graph discovery compared to STP walks, however, many other variants can be devised and fine tuned.

The immediate neighbors of a site  $s$  can be divided into the set  $V_{vv}$  of nodes that have already been visited and connected to  $s$  via visited edges, the set of nodes  $V_{vu}$  that have been already



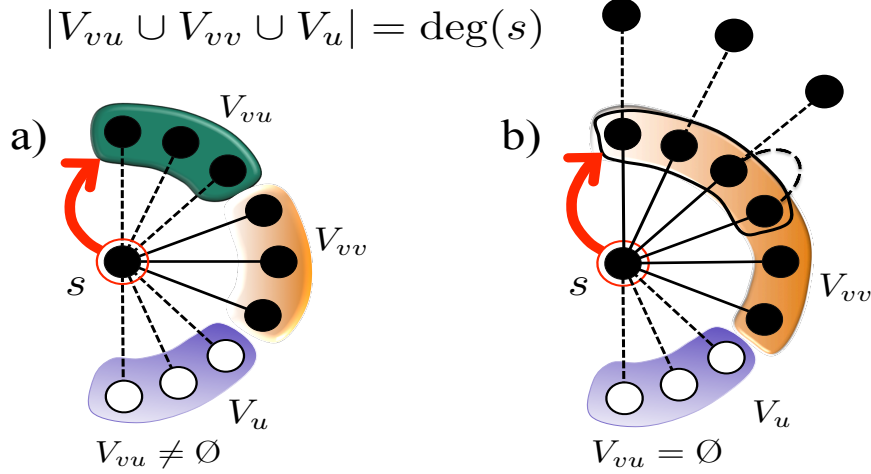


Figure 4: The Edge Explorer Model. (a) If there are already discovered neighbors  $V_{vu}$  of site  $s$  connecting through unvisited edges to  $s$ , the walker chooses one at random, uniformly, to move to. (b) If the visited neighbors are connected to  $s$  through visited edges only ( $V_{vu} = \emptyset$ ), the walker chooses one uniformly at random from those that have an unvisited connection to another visited node.

visited, but connected to  $s$  via unvisited edges and the set  $V_u$  of unvisited nodes (Fig. (4)). If  $V_{vu} \neq \emptyset$  (Fig. (4) a)), the walker steps to one of the nodes *from this set*, chosen uniformly at random. If, however,  $V_{vu} = \emptyset$  but  $V_{vv} \neq \emptyset$  (Fig. (4) b)), the walker chooses a node uniformly at random among the nodes within  $V_{vv}$  that have at least one connection via *unvisited* edges to other visited nodes. If no such nodes exist, then the walker chooses a node uniformly at random from  $V_u$ . While in general ATP walks do not lend to analytical treatment, all the properties of the EEM model on the complete graph  $K_N$  can be obtained exactly. Due to its rules, the walker always discovers *at least one edge* in two steps (only true on  $K_N$  !), which means that edge exploration happens linearly in time,  $\mu = 1$ . Assuming that it has discovered  $m - 1$  nodes, it will *not* discover a new node until it has discovered all the links amongst the  $m - 1$  nodes, the discovered graph becoming  $K_{m-1}$ . Then it adds the  $m$ -th node, discovering the remaining  $m - 1$  edges in  $\mathcal{O}(m)$  steps, thus finishing discovering all the nodes in  $\mathcal{O}(\sum m) = \mathcal{O}(N^2)$  steps. This means  $T_V = \mathcal{O}(N^2)$  and therefore  $\lambda = \ln N / \ln T_V = 1/2$ . Since on  $K_N$  the EEM walker cannot get lost in visited regions, the corresponding cross-over times and cover times are practically the same. Fig. (5)a) shows  $\langle S_n \rangle$  and  $\langle X_n \rangle$  with these predicted features, including  $\mu = 2\lambda = 1$ . On  $K_N$ , the edge exploration is optimal in the sense that for a given number of discovered nodes, it discovers the maximum number of edges possible up to that point, as shown in Fig. (5)b). Fig. (5)c-d) show the same for EEM on ER graphs. As discussed above, the  $\mu - \lambda$  slope difference increases with graph density, defined as  $\rho = \rho(G) = 2M/[N(N - 1)] \leq 1$ . On sparse graphs, however, the EEM can get trapped in visited regions if these regions are clusters/communities separated by bottlenecks from the rest of the graph. Within these regions the EEM walker performs a simple random walk before it escapes. For this reason, on low-density graphs the EEM is not

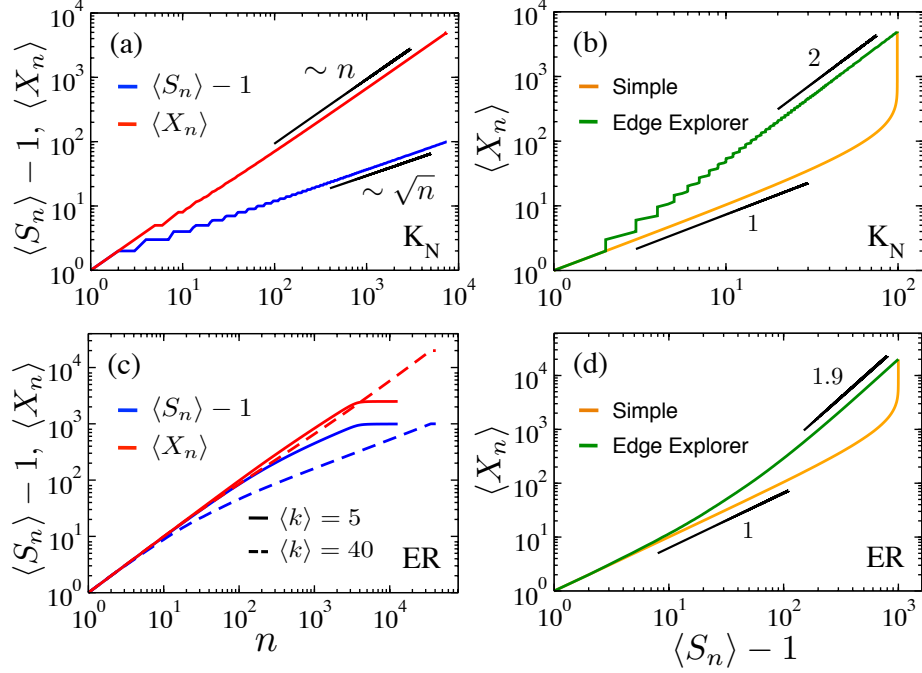


Figure 5: Site and edge exploration growth curves obtained for EEM on complete graphs ( $K_N$ ,  $N = 100$ , a) and b)), Erdős-Rényi graphs (ER,  $N = 1000$ , c) and d)).

necessarily the optimal explorer. To illustrate the graph discovering fidelity of the EEM, in Fig. (6) we compare the densities  $\bar{\rho}_n = 2\langle X_n \rangle / (\langle S_n \rangle (\langle S_n \rangle - 1))$  of the discovered graphs as function of time  $n$ , generated on the same network by the EEM and by the simple random walk (for  $K_N$  and ER). Clearly, the simple random walk greatly undershoots the true graph density (indicated by horizontal red line), corresponding to  $\langle X_n \rangle = \mathcal{O}(\langle S_n \rangle)$ , shown earlier, before it starts closing on the true value  $\rho(G)$ ; on the contrary, the EEM shows a systematic and rapid approach to  $\rho(G)$ .

## 6 Discussion

In summary, we have investigated properties of network discovery by walkers that follow edges (also called crawlers) in the most general setting. We have derived an exact expression for the average number of discovered edges  $\langle X_n \rangle$  (its generating function) as function of time for arbitrary graphs and STP walks. In particular, we have shown that for STP walkers both edge and node discovery follow the same scaling law on large networks, independently on the form of the stationary transition probabilities. Hence, the discovered network will be sparse (the number of discovered edges scaling linearly with that of the discovered nodes), presenting a strongly skewed structure compared to the underlying network's if the latter is not sparse,  $\nu > 1$ . Only after a cross-over time  $\sim \mathcal{O}(N)$ , will the edges become increasingly discovered, which in the case of large networks means unfeasibly large wait times, eliminating STP walks as a useful methodology for faithful net-

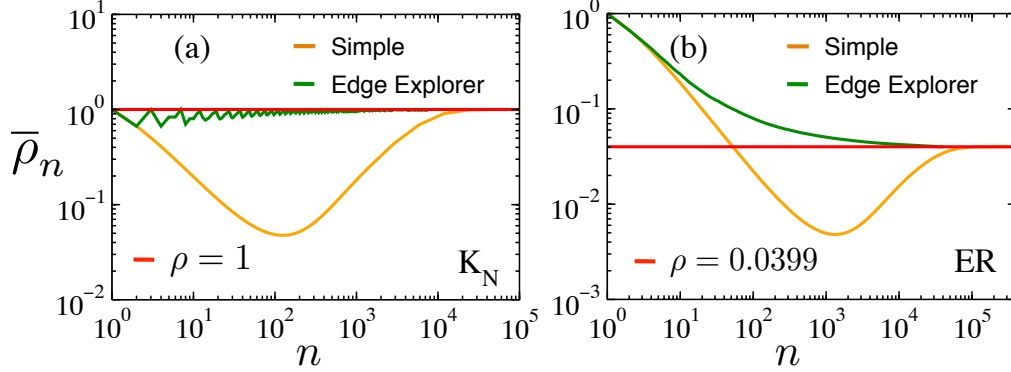


Figure 6: Comparison of the performance in graph discovery by the EEM adaptive walk (green) and by the simple random walk (orange), on a)  $K_N$  and b) ER graphs, by plotting the time evolution of the discovered graph's density. The red line is the true density. For a)  $N = 100$ , for b)  $N = 10^3$ ,  $\langle k \rangle = 40$ .

work discovery. Our results thus rigorously show that efficient/faithful discovery can only be done with adaptive walkers, whom use time/history dependent information for their transition probabilities (ATP). Visiting history information can be thought of as “pheromone” trails on the network, which the walker uses through its rules for stepping onto the next site [29]. There is a plethora of possible rules using past history, however, to keep memory requirements low (bounded) on a walker, the desirable rules are the ones that only use information from the local neighborhood of the walker. In this vein, we have introduced a simplistic adaptive walk, the Edge Explorer Model, which is greedily biased towards already visited regions within a 2-step neighborhood. We have shown that on dense graphs the EEM performs near optimally or optimally (on  $K_n$ ).

This project was supported in part by the Army Research Laboratory, ARL Cooperative Agreement Number W911NF-09-2-0053, HDTRA 201473-35045 and NSF BCS-0826958. The content of this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARL or the U.S. Government. The authors thank R.K.P Zia, B. Szymanski, M. Ercsey-Ravasz and S. Sreenivasan for useful discussions.

## A Derivation of $X(s_0, \xi)$

Consider a connected simple graph  $G(V, E)$  (there is a path along  $G$ 's edges between any two nodes), where  $V$  denotes the set of nodes and  $E$  the set of edges. In these pages we show the details of calculations for various exploration properties of (general) random walks on  $G$ , the only constraint being that the walk is restricted to move along existing edges (no long-range hops are allowed). We will rely heavily on the standard generating function technique [1, 30] and for that reason we briefly introduce related definitions and basic results. In particular, if  $A_j$  is an arbitrary

series, then its generating function  $A(\xi)$  with  $|\xi| < 1$  is defined as:

$$A(\xi) = \sum_{n=0}^{\infty} \xi^n A_n. \quad (10)$$

Knowing  $A(\xi)$ , the elements of the series  $A_n$  are recovered by inverting (10):

$$A_n = \frac{1}{2\pi i} \oint_{\Gamma} \frac{d\xi}{\xi^{n+1}} A(\xi), \quad (11)$$

where  $\Gamma$  is counterclockwise contour around  $\xi = 0$ . We are going to make use of the following expressions:

$$\sum_{j=0}^n A_j = \frac{1}{2\pi i} \oint_{\Gamma} \frac{d\xi}{\xi^{n+1}} \frac{1}{1-\xi} A(\xi), \quad \text{and} \quad \sum_{j=0}^{\infty} A_j = \lim_{\substack{\xi \rightarrow 1, \\ \xi < 1}} A(\xi) \equiv A(1^-). \quad (12)$$

On many occasions, the inversion integral in (11) cannot be performed analytically. However, we are usually interested in the long-time limit  $n \gg 1$  of the quantities, and for that we use the discrete Tauberian theorem which allows to estimate the leading order term, or the Darboux theorem (which can also produce terms beyond the leading order) [1, 2].

Let  $P_n(s|s_0)$  denote the probability of the walker being at site  $s$  on the  $n$ th step given that it started from site  $s_0$ , and let  $F_n(s|s_0)$  denote the probability of the walker visiting site  $s$  for the *first time* on the  $n$ -th step, given it started from  $s_0$ . The corresponding generating functions are  $P(s|s_0; \xi)$  and  $F(s|s_0; \xi)$  (Note that  $F(s|s_0; \xi) = \sum_{n=1}^{\infty} \xi^n F_n(s|s_0)$ ). Partition over the last step and partition over the first step give two useful recursion relations ( $r, s_0 \in V$ ):

$$P_{n+1}(r|s_0) = \sum_{r' \in V} p(r|r') P_n(r'|s_0), \quad P_{n+1}(r|s_0) = \sum_{r' \in V} P_n(r|r') p(r'|s_0). \quad (13)$$

In terms of generating functions:

$$P(r|s_0; \xi) = \delta_{rs_0} + \xi \sum_{r' \in V} p(r|r') P(r'|s_0; \xi), \quad (14)$$

$$P(r|s_0; \xi) = \delta_{rs_0} + \xi \sum_{r' \in V} P(r|r'; \xi) p(r'|s_0). \quad (15)$$

These identities can be used to derive a relationship between the site occupancy generating function and the first-passage time generating function, valid for *all* connected graphs, and *all* walks [2]:

$$F(s|s_0; \xi) = \frac{P(s|s_0; \xi) - \delta_{ss_0}}{P(s|s; \xi)}. \quad (16)$$

In the following we derive Eqs (5), (6) of the main paper. Let  $e = (s, s') \in E$  be an edge in  $G$ , and let  $F_n(e|s_0)$  denote the probability that the walker passes through  $e$  for the *first time* on

the  $n$ -th step, given it commenced from  $s_0 \in V$ . Let us denote by  $X_n(s_0)$  the number of distinct edges visited during an  $n$ -step walk that commenced from site  $s_0$ . We introduce the indicator  $\Gamma_n(s_0) \in \{0, 1\}$  for the number of virgin edges discovered on the  $n$ -th step:

$$\langle \Gamma_n(s_0) \rangle = \text{Prob}\{\Gamma_n(s_0) = 1\} = \sum_{e \in E} F_n(e|s_0) . \quad (17)$$

As convention we take  $\Gamma_0 = \langle \Gamma_0 \rangle = 0$ . Thus we can write:

$$X_n(s_0) = \sum_{j=0}^n \Gamma_j(s_0) , \quad \langle X_n(s_0) \rangle = \sum_{j=0}^n \langle \Gamma_j(s_0) \rangle , \quad (18)$$

The corresponding generating function is :

$$X(s_0; \xi) = \sum_{n=0}^{\infty} \xi^n \langle X_n(s_0) \rangle = \sum_{n=0}^{\infty} \xi^n \sum_{j=0}^n \langle \Gamma_j(s_0) \rangle = \frac{\Gamma(s_0; \xi)}{1 - \xi} , \quad (19)$$

where  $\Gamma(s_0; \xi)$  is the generating function for the indicator and we used the first identity in (12). From (17):

$$\Gamma(s_0; \xi) = \sum_{e \in E} \sum_{n=1}^{\infty} \xi^n F_n(e|s_0) = \sum_{e \in E} F(e|s_0; \xi). \quad (20)$$

We need to calculate the edge first passage probabilities  $F_n(e|s_0)$ , or their generating function  $F(e|s_0; \xi)$ . Clearly,  $F_n(e|s_0)$  contains all the paths commencing from  $s_0$  that *never crossed* edge  $e$  during the first  $n - 1$  steps, but they do so on the  $n$ -th step.

To calculate  $F(e|s_0; \xi)$ , we introduce an auxiliary node  $z$  placed on the edge  $e$ , as described in the main paper (Fig. 1c)), and consider the random walk on this extended graph  $G_z$ . The node set of this graph is  $V_z = V \cup \{z\}$  and the edge set  $E_z = E \setminus \{(s, s')\} \cup \{(s, z), (z, s')\}$ . Let  $P_n^\dagger(r|s_1)$  be the site occupation probability and  $F_n^\dagger(z|s_1)$  be the corresponding first passage time distribution for the random walk on  $G_z$ , and here  $r, s_1 \in V_z$ . Then it certainly holds that the first passage probability through edge  $e$  on the  $G$  graph is identical to the *site* first passage probability of the new walk to site  $z$  on the  $G_z$  graph:

$$F_n(e|s_0) = F_n^\dagger(z|s_0) , \quad s_0 \in V . \quad (21)$$

The corresponding generating function takes the form:

$$F(e|s_0; \xi) = F^\dagger(z|s_0; \xi) = \frac{P^\dagger(z|s_0; \xi)}{P^\dagger(z|z; \xi)} , \quad s_0 \neq z . \quad (22)$$

where we used the general identity (16) for the new walk on the new graph  $G_z$ .

In order to fully specify the new walk on  $G_z$  we need to define the corresponding single-step transition probabilities. The single-step probabilities away from the nodes  $s$  and  $s'$  of the new walk are identical to the old walk's, including those of getting to  $s$  and to  $s'$  from nodes other than  $s'$  and  $s$ , respectively. When considering steps from one  $(s(s'))$  to the other  $(s'(s))$ , the

single-step probabilities in the new walk are forbidden. Instead the walker has to step onto site  $z$  first, before it can step further to the other site ( $s$  or  $s'$ ). We also have to make sure that we have the same *probability flow* in the new walk as in the old one when moving from nodes  $s$  and  $s'$  towards  $z$  (in the old walk these would be along edge  $e$ ). From node  $z$  the walker can only step to  $s$  or  $s'$  with probabilities  $f$  and  $g$  respectively. The probabilities  $f$  and  $g$  are arbitrary, and the independence of the final expression for  $F(e|s_0; \xi)$  on these two variables serve as a good test for the correctness of our calculations. Eq (4) of the main paper provides the condensed form of the single-step transition probabilities for the new walk on  $G_z$ , and it combines the following cases:

$$p^\dagger(s|s') = p^\dagger(s'|s) = 0 \quad (23)$$

$$p^\dagger(s|z) = f, \quad p^\dagger(s'|z) = g, \quad p^\dagger(r|z) = 0 \quad \text{for } r \in V_z \setminus \{s, s'\} \quad (24)$$

$$p^\dagger(z|s') = p(s|s'), \quad p^\dagger(z|s) = p(s'|s), \quad p^\dagger(z|r') = 0 \quad \text{for } r' \in V_z \setminus \{s, s'\} \quad (25)$$

$$p^\dagger(r|r') = p(r|r') + q(r|r'), \quad q(r|r') = -(\delta_{r's'}\delta_{rs} + \delta_{r's}\delta_{rs'})p(r|r'), \quad r, r' \in V. \quad (26)$$

The pseudo-node  $z$  uniquely characterizes the edge  $e = (s, s') \in E$ . To remind us about this identification, we will write  $z$  instead of  $e$ , in the remainder. From (22) and (20) it follows that:

$$\Gamma(s_0; \xi) = \sum_{z \in E} \frac{P^\dagger(z|s_0; \xi)}{P^\dagger(z|z; \xi)}, \quad s_0 \in V. \quad (27)$$

The sum is over all the edges of the *original* graph  $G$ . Thus we need to compute the site occupation probabilities for the new walk on the extended graph. The relationships based on partition over the last step (14) and first step (15), hold for the new walk as well:

$$P^\dagger(r|s_1; \xi) = \delta_{rs_1} + \xi \sum_{r' \in V_z} p^\dagger(r|r')P^\dagger(r'|s_1; \xi), \quad r, s_1 \in V_z \quad (28)$$

$$P^\dagger(r|s_1; \xi) = \delta_{rs_1} + \xi \sum_{r' \in V_z} P^\dagger(r|r'; \xi)p^\dagger(r'|s_1), \quad r, s_1 \in V_z. \quad (29)$$

From (28) with  $r = z$  and  $s_1 = s_0$  and using (25) one obtains:

$$P^\dagger(z|s_0; \xi) = \xi p(s|s')P^\dagger(s'|s_0; \xi) + \xi p(s'|s)P^\dagger(s|s_0; \xi), \quad s_0 \in V. \quad (30)$$

Eq (29) with  $r = s_1 = z$  and (24) yields:

$$P^\dagger(z|z; \xi) = 1 + \xi f P^\dagger(z|s; \xi) + \xi g P^\dagger(z|s'; \xi). \quad (31)$$

Since (30) is valid for any  $s_0 \in V$ , we first make  $s_0 \mapsto s$  and then the  $s_0 \mapsto s'$  substitutions to get :

$$P^\dagger(z|s; \xi) = \xi p(s|s')P^\dagger(s'|s; \xi) + \xi p(s'|s)P^\dagger(s|s; \xi) \quad (32)$$

$$P^\dagger(z|s'; \xi) = \xi p(s|s')P^\dagger(s'|s'; \xi) + \xi p(s'|s)P^\dagger(s|s'; \xi) \quad (33)$$

Inserting these into (31) one obtains:

$$\begin{aligned} P^\dagger(z|z; \xi) &= 1 + \xi^2 p(s|s') \left[ f P^\dagger(s'|s; \xi) + g P^\dagger(s'|s'; \xi) \right] + \\ &\quad + \xi^2 p(s'|s) \left[ f P^\dagger(s|s; \xi) + g P^\dagger(s|s'; \xi) \right] \end{aligned} \quad (34)$$

Eq (34) shows that we need to calculate the site occupation probability generating function  $P^\dagger(r|s_0; \xi)$  for  $r, s_0 \in V$ . In order to derive  $P^\dagger(r|s_0; \xi)$  for  $r, s_0 \in V$  we first make the replacements  $r' \mapsto r''$ ,  $r \mapsto r'$ ,  $s_1 \mapsto s_0 \in V$  in (29) to obtain:

$$P^\dagger(r'|s_0; \xi) = \delta_{r's_0} + \xi \sum_{r'' \in V_z} p^\dagger(r'|r'') P^\dagger(r''|s_0; \xi), \quad (35)$$

With the help of (26) we then find:

$$P^\dagger(r'|s_0; \xi) = \delta_{r's_0} + \xi \sum_{r'' \in V} \{p(r'|r'') + q(r'|r'')\} P^\dagger(r''|s_0; \xi) + \xi p^\dagger(r'|z; \xi) P^\dagger(z|s_0; \xi)$$

or after rearranging

$$\begin{aligned} P^\dagger(r'|s_0; \xi) - \xi \sum_{r'' \in V} p(r'|r'') P^\dagger(r''|s_0; \xi) &= \delta_{r's_0} + \xi \sum_{r'' \in V} q(r'|r'') P^\dagger(r''|s_0; \xi) \\ &+ \xi p^\dagger(r'|z; \xi) P^\dagger(z|s_0; \xi). \end{aligned} \quad (36)$$

After multiplying both sides with  $P(r|r'; \xi)$ ,  $r \in V$  and summing both sides over  $r' \in V$ , the above equation takes the form of

$$\begin{aligned} \sum_{r' \in V} P(r|r'; \xi) P^\dagger(r'|s_0; \xi) - \sum_{r'' \in V} P^\dagger(r''|s_0; \xi) \xi \sum_{r' \in V} p(r'|r'') P(r|r'; \xi) &= P(r|s_0; \xi) + \\ &+ \sum_{r'' \in V} P^\dagger(r''|s_0; \xi) \xi \sum_{r' \in V} P(r|r'; \xi) q(r'|r'') + \\ &+ \xi [fP(r|s; \xi) + gP(r|s'; \xi)] P^\dagger(z|s_0; \xi). \end{aligned} \quad (37)$$

Next we calculate the left hand side of (37) by using (15) to write:

$$\xi \sum_{r' \in V} p(r'|r'') P(r|r'; \xi) = P(r|r''; \xi) - \delta_{rr''}$$

When this is inserted into the lhs of (37), the sums with dagger terms cancel and one just simply obtains  $P^\dagger(r|s_0; \xi)$ . The sums on the right hand side can be written in a simpler form after introducing the notation:

$$A(r|r''; \xi) = \xi \sum_{r' \in V} P(r|r'; \xi) q(r'|r''). \quad (38)$$

Thus, eq (37) assumes the expression:

$$\begin{aligned} P^\dagger(r|s_0; \xi) &= P(r|s_0; \xi) + \xi [fP(r|s; \xi) + gP(r|s'; \xi)] P^\dagger(z|s_0; \xi) \\ &+ \sum_{r'' \in V} A(r|r''; \xi) P^\dagger(r''|s_0; \xi). \end{aligned} \quad (39)$$

Using (26) we find:

$$A(r|r''; \xi) = -\xi P(r|s; \xi) p(s|r'') \delta_{r''s'} - \xi P(r|s'; \xi) p(s'|r'') \delta_{r''s} \quad (40)$$

Inserting this in (39) one obtains:

$$P^\dagger(r|s_0; \xi) = P(r|s_0; \xi) - \xi P(r|s; \xi) p(s|s') P^\dagger(s'|s_0; \xi) - \xi P(r|s'; \xi) p(s'|s) P^\dagger(s|s_0; \xi) + \xi [f P(r|s; \xi) + g P(r|s'; \xi)] P^\dagger(z|s_0; \xi), \quad r \in V. \quad (41)$$

Replacing  $r \mapsto s$  and then  $r \mapsto s'$  in the equation above, yields:

$$\begin{cases} a_{11} P^\dagger(s|s_0; \xi) + a_{12} P^\dagger(s'|s_0; \xi) + a_{13} P^\dagger(z|s_0; \xi) = P(s|s_0; \xi) \\ a_{21} P^\dagger(s|s_0; \xi) + a_{22} P^\dagger(s'|s_0; \xi) + a_{23} P^\dagger(z|s_0; \xi) = P(s'|s_0; \xi) \end{cases} \quad (42)$$

where:

$$\begin{aligned} a_{11} &= 1 + \xi p(s'|s) P(s|s'; \xi) & a_{21} &= \xi p(s'|s) P(s'|s'; \xi) \\ a_{12} &= \xi p(s|s') P(s|s; \xi) & a_{22} &= 1 + \xi p(s|s') P(s'|s; \xi) \\ a_{13} &= -\xi [f P(s|s; \xi) + g P(s|s'; \xi)] & a_{23} &= -\xi [f P(s'|s; \xi) + g P(s'|s'; \xi)] \end{aligned} \quad (43)$$

System (42) needs a third equation, to solve for  $\{P^\dagger(s|s_0; \xi), P^\dagger(s'|s_0; \xi), P^\dagger(z|s_0; \xi)\}$ . The third equation is just (30):

$$a_{31} P^\dagger(s|s_0; \xi) + a_{32} P^\dagger(s'|s_0; \xi) + a_{33} P^\dagger(z|s_0; \xi) = 0 \quad (44)$$

with:

$$a_{31} = \xi p(s'|s), \quad a_{32} = \xi p(s|s'), \quad a_{33} = -1 \quad (45)$$

Thus, if we introduce the column vectors:

$$[P^\dagger](s, s', z|s_0; \xi) \equiv \begin{bmatrix} P^\dagger(s|s_0; \xi) \\ P^\dagger(s'|s_0; \xi) \\ P^\dagger(z|s_0; \xi) \end{bmatrix}, \quad [P](s, s'|s_0; \xi) \equiv \begin{bmatrix} P(s|s_0; \xi) \\ P(s'|s_0; \xi) \\ 0 \end{bmatrix}, \quad (46)$$

and denote by  $\mathbf{A}$  the  $3 \times 3$  matrix with elements defined above by (43) and (45), then the linear equation to be solved is simply:

$$\mathbf{A} [P^\dagger](s, s', z|s_0; \xi) = [P](s, s'|s_0; \xi). \quad (47)$$

Assuming that  $\mathbf{A}$  is invertible, the solution is

$$[P^\dagger](s, s', z|s_0; \xi) = \mathbf{A}^{-1} [P](s, s'|s_0; \xi) \quad (48)$$

The matrix explicitly looks as follows:

$$\mathbf{A} = \begin{bmatrix} 1 + \xi \alpha a & \xi \beta b & -\xi(ga + fb) \\ \xi \alpha c & 1 + \xi \beta d & -\xi(gc + fd) \\ \xi \alpha & \xi \beta & -1 \end{bmatrix} \quad (49)$$

where we introduced the shorthand notations:

$$\alpha = p(s'|s), \quad \beta = p(s|s'), \quad (50)$$

$$a = a(\xi) = P(s|s'; \xi), \quad b = b(\xi) = P(s|s; \xi), \quad (51)$$

$$c = c(\xi) = P(s'|s'; \xi), \quad d = d(\xi) = P(s'|s; \xi) \quad (52)$$



The inverse of matrix  $\mathbf{A}$  is just

$$\mathbf{A}^{-1} = \frac{1}{D} \begin{bmatrix} -1 - \beta(d - v\xi)\xi & \beta(b - u\xi)\xi & (u + t\beta g\xi)\xi \\ \alpha(c - v\xi)\xi & -1 - \alpha(a - u\xi)\xi & (v + t\alpha f\xi)\xi \\ -\alpha(1 + \bar{v}\beta\xi)\xi & -\beta(1 + \bar{u}\alpha\xi)\xi & 1 + (w + t\alpha\beta\xi)\xi \end{bmatrix} \quad (53)$$

where the determinant is:

$$D = -1 - w\xi + (u\alpha + v\beta - t\alpha\beta)\xi^2 + t\alpha\beta(f + g)\xi^3, \quad (54)$$

and we introduced the notations:

$$\begin{aligned} u &= ga + fb & \bar{u} &= a - b \\ v &= fd + gc & \bar{v} &= d - c \\ t &= ad - bc & w &= a\alpha + d\beta. \end{aligned} \quad (55)$$

The final solutions can be easily read from (48). Note that these are now expressed solely in terms of the generating functions for the site occupation probabilities of the old walk on the old graph  $G$ ! In particular for  $P^\dagger(z|s_0; \xi)$  we get:

$$P^\dagger(z|s_0; \xi) = -\frac{\alpha\xi}{D}(1 + \bar{v}\beta\xi)P(s|s_0; \xi) - \frac{\beta\xi}{D}(1 + \bar{u}\alpha\xi)P(s'|s_0; \xi). \quad (56)$$

In order to obtain the site occupation probability generating functions  $P^\dagger(s|s; \xi)$ ,  $P^\dagger(s'|s; \xi)$ ,  $P^\dagger(s|s'; \xi)$  and  $P^\dagger(s'|s'; \xi)$  as required by the r.h.s. of (34) we merely substitute  $s$  and  $s'$ , respectively for  $s_0$  in the expressions for the solutions. We thus obtain:

$$P^\dagger(s|s; \xi) = -\frac{1}{D}(b + g\beta t\xi^2), \quad P^\dagger(s|s'; \xi) = -\frac{1}{D}(a + t\beta\xi - t\beta f\xi^2) \quad (57)$$

$$P^\dagger(s'|s; \xi) = -\frac{1}{D}(d + t\alpha\xi - t\alpha g\xi^2), \quad P^\dagger(s'|s'; \xi) = -\frac{1}{D}(c + t\alpha f\xi^2). \quad (58)$$

Inserting these into (34):

$$P^\dagger(z|z; \xi) = -\frac{1}{D}(1 + w\xi + t\alpha\beta\xi^2). \quad (59)$$

Next, from Eqs. (56) (59) and (27) we find:

$$\Gamma(s_0; \xi) = \sum_{z \in E} \frac{\alpha\xi[1 + (d - c)\beta\xi]P(s|s_0; \xi) + \beta\xi[1 + (a - b)\alpha\xi]P(s'|s_0; \xi)}{1 + (a\alpha + d\beta)\xi + (ad - bc)\alpha\beta\xi^2}. \quad (60)$$

Note that this expression is independent on the variables  $f$  and  $g$ , as anticipated! Now using (19), the generating function for the average number of discovered distinct edges becomes:

$$X(\xi) = \frac{\xi}{2(1 - \xi)} \sum_{s, s'} \frac{\alpha[1 + (d - c)\beta\xi]P(s|s_0; \xi) + \beta[1 + (a - b)\alpha\xi]P(s'|s_0; \xi)}{1 + (a\alpha + d\beta)\xi + (ad - bc)\alpha\beta\xi^2}. \quad (61)$$

Taking a closer look at this expression one observes that only those  $(s, s')$  pairs will contribute in the sum which are neighbors on  $G$  (since  $\alpha$  and  $\beta$  are zero for transitions along non-edges). After adding the sum to itself then interchanging the dummy variables  $s, s'$  in one of the sums, we finally obtain eqs (5),(6) of the main paper:

$$X(s_0; \xi) = \frac{\xi}{1-\xi} \sum_s \overline{W}(s; \xi) P(s|s_0; \xi), \quad (62)$$

with

$$\overline{W}(s; \xi) = \sum_{s'} \alpha \frac{1 + (d-c)\beta\xi}{1 + (a\alpha + d\beta)\xi + (ad-bc)\alpha\beta\xi^2}. \quad (63)$$

## B Properties of $X(s_0; \xi)$

Next we show a number of properties of the weight function  $\overline{W}(s; \xi)$ . Before we do that, however, we need to establish a number of fundamental inequalities involving the generating functions (51), (52). Let us use the temporary notation (where all quantities are understood implicitly in the  $\xi \rightarrow 1^-$  limit,  $\xi \in \mathbb{R}$ ):

$$\Delta \equiv 1 + a\alpha + d\beta + (ad-bc)\alpha\beta \quad (64)$$

The denominator  $\Delta$  of (63) appears in the numerator of  $P^\dagger(z|z; \xi)$  given by (59) and thus:

$$P^\dagger(z|z; \xi) = -\frac{\Delta}{D}. \quad (65)$$

Since the  $P^\dagger$ -s in (57-59) are all generating functions for random walk probabilities, (that is, they are power-series with positive coefficients) they are all positive (for  $\xi \in \mathbb{R}$ ,  $\xi \rightarrow 1^-$ ). Hence if we show that the determinant  $D \leq 0$ , then from the positivity of  $P^\dagger(z|z; \xi)$  it follows that  $\Delta \geq 0$ . From (54) in the  $\xi \rightarrow 1^-$  limit:  $-D = 1 + w - (u\alpha + v\beta)$ , where we used the normalization  $f + g = 1$ . Let us examine the sign of this expression. Using the definitions from (55), we obtain:

$$-D = 1 + a\alpha(1-g) + d\beta(1-f) - \alpha fb - \beta gc. \quad (66)$$

Recall, that  $f + g = 1$ , with  $f$  and  $g$  being probabilities chosen arbitrarily. Then  $D$  can be rewritten as

$$-D = 1 + f\alpha(a-b) + g\beta(d-c). \quad (67)$$

Determinant  $D$  also appears in the expressions of the site occupancy generating functions  $P^\dagger(s|s; \xi)$  and  $P^\dagger(s|s'; \xi)$ , in (57). In the  $\xi \rightarrow 1^-$  limit, these expressions take the form of

$$P^\dagger(s|s; 1^-) = -\frac{1}{D}(b + g\beta t), \quad (68)$$

$$P^\dagger(s|s'; 1^-) = -\frac{1}{D}(a + g\beta t). \quad (69)$$

The sum  $cP^\dagger(s|s; 1^-) + dP^\dagger(s|s'; 1^-)$  is always non-negative ( $c, d \geq 0$ ), implying

$$cP^\dagger(s|s; 1^-) + dP^\dagger(s|s'; 1^-) = -\frac{1}{D} [ad + bc + t\beta g(c + d)] \geq 0. \quad (70)$$

As  $a, b, c, d \geq 0$  (for  $\xi \rightarrow 1^-$ ,  $\xi \in \mathbb{R}$ ),  $0 \leq \beta \leq 1$ , and  $g \in [0, 1]$  is an arbitrarily chosen transition probability, for any given  $\xi$ ,  $g$  can always be chosen to be small enough, such that  $ad + bc + t\beta g(c + d) \geq 0$ , independently on the sign of  $t$ . (Note that  $a, b, c, d$  are independent on  $g$  or  $f$ ). Thus, for small  $g$  values,  $D \leq 0$ , and based on (65) this implies that  $\Delta \geq 0$ . Since  $\Delta$  is independent of  $f$  and  $g$ , this also implies that for arbitrary  $f$  ( $g = 1 - f$ ), we have:

$$-D = 1 + f\alpha(a - b) + g\beta(d - c) \geq 0, \quad \text{as } \xi \rightarrow 1^- \quad (71)$$

$$\Delta = 1 + (a\alpha + d\beta) + (ad - bc)\alpha\beta \geq 0, \quad \text{as } \xi \rightarrow 1^- \quad (72)$$

Hence, if we choose  $f = 1, g = 0$  in (71), followed by  $f = 0, g = 1$  we obtain:

$$1 + \alpha(a - b) \geq 0, \quad (73)$$

$$1 + \beta(d - c) \geq 0, \quad (74)$$

as  $\xi \rightarrow 1^-$ . Observe that (74) is also obtained by switching  $s, s'$ , to  $s'$  and  $s$  respectively, in (73) (which holds for any  $s$  and  $s'$ ). Using (16), and the fact that  $F(s|s_0; \xi) \leq 1$  even at  $|\xi| = 1$  ( $F(s|s_0; 1) = R(s|s_0)$  is the *probability* that  $s$  is ever reached from  $s_0$ ), after replacing  $s_0$  by  $s' \neq s$ , we find that in the limit  $\xi \rightarrow 1^-$

$$a - b \leq 0. \quad (75)$$

However, although  $a \leq b$ , the difference  $a - b$  cannot be arbitrarily negative, as shown in (73):  $a - b \geq -\alpha^{-1}$ . Similarly, it holds:

$$d - c \leq 0. \quad (76)$$

An immediate consequence of these inequalities, is that in the case of recurrent walks, where both  $a$  and  $b$  (and similarly,  $c$  and  $d$ ) diverge as  $\xi \rightarrow 1^-$ , their difference is nevertheless bounded. From (74) and (72) it follows that every term in the expression of  $\overline{W}(s; 1^-)$  is non-negative, and thus  $\overline{W}(s; 1^-) \geq 0$ .

Next we show that  $\overline{W}(s; 1^-) < \infty$  (finite). Recall, that  $1 \leq X_n \leq n$ , that is, the number of edges cannot grow faster than linearly with time (at most one new edge can be discovered in a time step). This inequality written in terms of generating functions becomes:

$$\frac{1}{1 - \xi} \leq X(\xi) \leq \frac{\xi}{(1 - \xi)^2}, \quad \xi \in \mathbb{R}, \quad \xi > 0. \quad (77)$$

Let us assume that on the contrary,  $\overline{W}(s; \xi) \rightarrow \infty$  as  $\xi \rightarrow 1^-$ . This means that for *any arbitrarily large constant*  $C$ , there is a  $\xi_0 \in \mathbb{R}$ ,  $\xi_0 < 1$ , such that  $\overline{W}(s; \xi) > C$  for all  $\xi > \xi_0$ ,  $\xi \rightarrow 1^-$ . Thus:

$$X(\xi) = \frac{\xi}{1 - \xi} \sum_s \overline{W}(s; \xi) P(s|s_0; \xi) > \frac{C\xi}{1 - \xi} \sum_s P(s|s_0; \xi) = \frac{C\xi}{(1 - \xi)^2}, \quad (78)$$

where in the last step we used the identity  $\sum_s P(s|s_0) = (1 - \xi)^{-1}$ , which is a direct consequence of the normalization condition  $\sum_s P_n(s|s_0) = 1$ ). Since  $C$  is arbitrarily large, it can certainly be chosen such that  $C > 1$ , and thus (78) will be contradicting (77).

Let us write (63) in the form:

$$\overline{W}(s; \xi) = \sum_{s'} \alpha \frac{1 + (d - c)\beta\xi}{1 + (a\alpha + d\beta)\xi + (ad - bc)\alpha\beta\xi^2} = \sum_{s'} \alpha \theta(s, s', \xi). \quad (79)$$

Functions  $a$  and  $d$  can be expressed in terms of the first-passage time generating function by using relation (16):  $a = F(s|s'; \xi)P(s|s; \xi) = \eta b$ ,  $d = F(s'|s; \xi)P(s'|s'; \xi) = \eta' c$ , where we have introduced the notations  $\eta = F(s|s'; \xi) \leq 1$  and  $\eta' = F(s'|s; \xi) \leq 1$ . In the  $\xi \rightarrow 1^-$  limit, we obtain for  $\theta(s, s', 1^-)$

$$\theta(s, s', 1^-) = \frac{1 - (1 - \eta')c\beta}{1 + b\alpha\eta + c\beta\eta' - bc(1 - \eta\eta')\alpha\beta}. \quad (80)$$

The denominator can be written as:  $1 + b\alpha\eta + c\beta\eta' - bc(1 - \eta\eta')\alpha\beta = [1 - (1 - \eta')c\beta][1 + (1 + \eta)b\alpha] - b\alpha(1 + c\beta\eta') + c\beta(1 + b\alpha\eta)$ . Let  $u' = 1 - (1 - \eta')c\beta$  and  $u = 1 - (1 - \eta)b\alpha$ . Since  $\eta'c = d$  and  $\eta b = a$ , we have  $u' = 1 + \beta(d - c)$  and  $u = 1 + \alpha(a - b)$ . Thus, from (73-74) it follows that  $u' \geq 0$  and  $u \geq 0$ . Moreover, we can write  $1 - \eta' = (1 - u')/(c\beta)$ ,  $1 - \eta = (1 - u)/(b\alpha)$ . Since  $1 - \eta' \geq 0$  and  $1 - \eta \geq 0$  (see above) and as both  $c\beta$  and  $b\alpha$  are positive it follows that one must have:

$$0 \leq u' \leq 1, \quad 0 \leq u \leq 1. \quad (81)$$

Using the  $u'$  notation, equation Eq. (80) is written as:

$$\theta(s, s', 1^-) = \frac{u'}{u'[1 + (1 + \eta)b\alpha] + c\beta(1 + b\alpha\eta) - b\alpha(1 + c\beta\eta')}. \quad (82)$$

In the denominator we then use  $1 + c\beta\eta' = u' + c\beta$ . After the cancelations we find:

$$\theta(s, s', 1^-) = \frac{u'}{u'(1 + b\alpha\eta) + c\beta(1 + b\alpha\eta) - bc\alpha\beta} = \frac{u'}{uu' + u'b\alpha + uc\beta}. \quad (83)$$

The last equality was obtained after using  $1 + b\alpha\eta = u + b\alpha$ . Hence:

$$\overline{W}(s; 1^-) = \frac{1}{b} \sum_{s'} \alpha \frac{u'}{u'\alpha + u\varphi\beta + uu'/b}, \quad (84)$$

where  $\varphi = c/b = P(s'|s'; 1^-)/P(s|s; 1^-)$ . Taking into account (81), the terms in the sum are all positive and finite (even if  $b \rightarrow \infty$ ), that is there is  $C > 0$  constant such that:

$$\overline{W}(s; 1^-) < \frac{1}{b} \sum_{s'} \alpha C = \frac{1}{b} C. \quad (85)$$

(Recall, that due to normalization  $\sum_{s'} \alpha = \sum_{s'} p(s'|s) = 1$ .) This implies from (62) that in the limit  $\xi \rightarrow 1^-$  ( $\xi$  is very close to 1 but not quite 1):

$$X(s_0; \xi) < \frac{1}{1 - \xi} \sum_s C \frac{P(s|s_0; \xi)}{b} = CS(s_0; \xi) \quad (86)$$

Clearly, after the first loop was made by the walker (which happens after long enough times, otherwise we have  $X_n = S_n - 1$  identically, in which case trivially  $\mu = \lambda$ ), we have  $S_n < X_n$  and hence:

$$S(s_0; \xi) < X(s_0; \xi) < CS(s_0; \xi), \quad \text{as } \xi \rightarrow 1^-, \quad (87)$$

showing that the leading order behavior for  $\langle X_n \rangle$  and  $\langle S_n \rangle$  are identical, i.e.,  $\mu = \lambda$ .

**Discussion.** Recall, that  $\langle S_n \rangle \sim n^\lambda$  ( $\langle X_n \rangle \sim n^\mu$ ) means  $\langle S_n \rangle \simeq n^\lambda L(n)$  ( $\langle X_n \rangle \simeq n^\mu F(n)$ ) as  $n \rightarrow \infty$ , where  $L(n), F(n)$  are slowly varying functions, that is  $L(\zeta x)/L(x) \rightarrow 1$  ( $F(\zeta' x)/F(x) \rightarrow 1$ ) when  $x \rightarrow \infty$ , for any  $\zeta, \zeta' > 0$ .

- i) When  $\mu = \lambda \neq 0$ ,  $\langle S_n \rangle$  and  $\langle X_n \rangle$  obey the same scaling laws, however for early times, the corrections  $L(n)$  and  $F(n)$  can be dominant. This results in a slight deviation between the two curves, e.g., Sierpinski gasket, square lattice, as seen in the main article. For large graphs, in the  $n \rightarrow \infty$  limit, these corrections diminish.
- ii) When  $\mu = \lambda = 0$ , the growth of  $\langle S_n \rangle$  is characterized by the function  $L(n)$ , while the growth of  $\langle X_n \rangle$  is dictated by the  $F(n)$ , which are not necessarily the same. This case is shown in Fig. (7) for the scale-free BA model. For simple random walks,  $\langle S_n \rangle$  and  $\langle X_n \rangle$  grow at the same rate, while for a walk that is biased towards high degree nodes, after a short linear growth region they slowly grow following different curves till they saturate. For the biased STP walk towards higher degrees we took  $p(s'|s) = k_s \left( \sum_{s'' \in \langle s \rangle} k_{s''} \right)^{-1}$  where  $k_s$  is the degree of node  $s$  and  $\langle s \rangle$  denotes the set of graph neighbors of  $s$ .

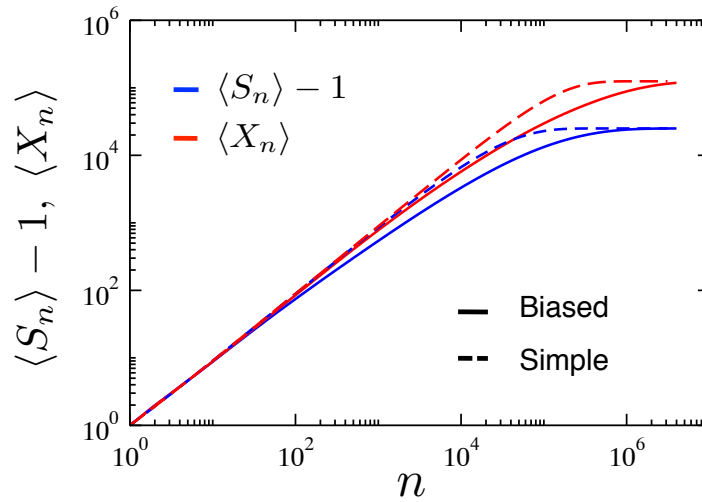


Figure 7: Comparison of the average number of discovered nodes (blue lines) and edges (red lines) on scale-free BA model ( $N = 25000$ ,  $\langle k \rangle = 10$ ), for simple random walk (dashed lines) and for a walk that is biased towards high degree nodes (solid lines).

## References

- [1] Barry D. Hughes. *Random Walks and Random Environments*, volume 1: Random Walks. Oxford U. P., 1995.
- [2] William Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, 2 edition, 1971.
- [3] David J. Aldous and James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs*. 1999.
- [4] Santosh Vempala. Geometric random walks: A survey. *Comb. Comp. Geom.*, 52:573–612, 2005.
- [5] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Mark Najork. In *Proc. of the 8th Intern. World Wide Web Conf., Toronto, Canada*, page 213. Elsevier Science, May 1999.
- [6] Petter Holme. Congestion and centrality in traffic flow on complex networks. *Adv. Comp. Sys.*, 6:163–176, 2003.
- [7] Jae Dong Noh and Heiko Rieger. Random walks on complex networks. *Phys. Rev. Lett.*, 92(11):118701, 2004.
- [8] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, 2001.
- [9] Mark E.J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, 2002.
- [10] D. Volchenkov, L. Volchenkova, and Ph. Blanchard. Epidemic spreading in a variety of scale free networks. *Phys. Rev. E*, 66:046137, 2002.
- [11] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, 2001.
- [12] Steven H. Strogatz. Exploring complex networks. *Nature*, 410:268, 2001.
- [13] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.
- [14] A. Dvoretzky and P. Erdős. In J. Neyman, editor, *Proc. of the 2nd Berkeley Symp. on Math. Stat. Prob.*, pages 353–367. U. California Press, Berkeley, 1951.
- [15] Elliott W. Montroll and George H. Weiss. Random walks on lattices. II. *J. Math. Phys.*, 6:167–181, 1965.
- [16] R. Burioni and D. Cassi. Random walks on graphs: idea, techniques and results. *J. Phys. A: Math. Gen.*, 38:R45–R78, 2005.

- [17] L. Lovász. Random walks on graphs: a survey. *Combinatorics, Paul Erdős is Eighty*, 2:1–46, 1993.
- [18] Jani Lahtinen, János Kertész, and Kimmo Kaski. Scaling of random spreading in small world networks. *Phys. Rev. E*, 64:057105, 2001.
- [19] E. Almaas, R. V. Kulkarni, and D. Stroud. Scaling properties of random walks on small-world networks. *Phys. Rev. E*, 68:056105, 2003.
- [20] Greg Barnes and Uriel Feige. Short random walks on graphs. *Annual ACM Symposium on Theory of Computing*, pages 728–737, 1993.
- [21] A. Ramezanzpour. Intermittent exploration on a scale-free network. *Europhys. Lett.*, 77:60004, 2007.
- [22] K. L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, Berlin, 1960.
- [23] P. Erdős and A. Rényi. On random graphs I. *Publ. Math.*, 6:290–297, 1959.
- [24] Jesper Dall and Michael Christensen. Random geometric graphs. *Phys. Rev. E*, 66:016121, 2002.
- [25] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
- [26] Erzsébet Ravasz, A. L. Somera, D. A. Mongru, Zoltán N. Oltvai, and Albert-László Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551, 2002.
- [27] Lazaros K. Gallos, Chaoming Song, Shlomo Havlin, and Hernán A. Makse. Scaling theory of transport in complex biological networks. *PNAS*, 104(19):7746–7751, May 2007.
- [28] J. C. Angles d’Auriac, A. Benoit, and R. Rammal. Random walk on fractals: numerical studies in two dimensions. *Jour. Phys. A*, 16:4039–4051, 1983.
- [29] Sven Koenig, Boleslaw Szymanski, and Yaxin Liu. Efficient and inefficient ant coverage methods. *Annals Math. & Art. Int.*, 31:41–76, 2001.
- [30] H. S. Wilf. *generatingfunctionology*. A. K. Peters LTD, 3rd edition, 2005.